More for Less? Comparing small area estimation, spatial microsimulation, and mass imputation

Stephen Haslett[1], Geoffrey Jones[2], Alasdair Noble[3], Dimitris Ballas[4]

[1, 2, 3]Institute of Fundamental Sciences – Statistics, Massey University, PO Box 11222
Palmerston North, New Zealand
[4] Department of Geography, University of Sheffield, S10 2TN, United Kingdom

Abstract
Combining data sources is often seen as a panacea, having the potential to produce more to produce cost-effective, accurate, fine-level statistics for a lower cost. This paper clarifies conditions under which Official Statistics data sources, particularly surveys and censuses or surveys and administrative sources, should and should not be combined using statistical models based on mass imputation, spatial microsimulation, and small area and domain estimation. The theoretical links between these three techniques are explored. The wider research from which this paper is a report considers the relevant literature in depth, further develops existing statistical methods, considers their application in principle to set of case studies in sociology, economics, and business, and provides guidelines for use of the three techniques based on this research.

Key Words: spatial microsimulation ; small area estimation ; mass imputation ; combining data

1. Introduction

Internationally across Official Statistics, there is wide discussion and increased usage of techniques which combine survey with census or administrative data to reduce overall costs and/or provide more detailed, finer-level statistics. There is also increasing awareness that emphasis in Official Statistics needs to continue toward supplementing data collection with better data utilisation. However the underlying statistical theory to do this well requires further research, especially for modelling methods that combine data as anonymously as possible to limit confidentiality concerns.

Mass imputation, spatial microsimulation, and small area estimation (including small domain estimation) are three statistical methods for extending usage of survey and census data. Although their research literature is essentially separate, the three techniques have underlying similarities. The nature and extent of those similarities, along with the important differences, are here linked to guidelines for their proper use. A more extensive outline of the research is given in Haslett, Jones, Noble and Ballas (2010).

The principal aim of this research project has been to explore how best to combine survey and census data, or survey and administrative data, using sound statistical models to produce finer-level statistics for variables collected only by sample survey, without formally linking individual records and raising confidentiality issues. Essentially, the methods work by imputing or predicting variables of interest that are not collected or are partially missing in the larger dataset. Prediction is usually via models that use variables common to both datasets. The fitted model can then provide multiple predictions for all missing census observations, which when combined can give not only area or subpopulation estimates but also with estimates of accuracy conditional on the model being correct. The statistical techniques considered include mass imputation, spatial microsimulation, and certain types of small area and small domain methods.

The research has allowed development of guidelines on when, in principle, methods that combine data from different sources can and cannot assist Official Statistics agencies to reduce respondent load and cost, and to improve accuracy of existing surveys by use of supplementary data sources. General comment is also possible on how surveys best be designed in future to integrate with census and administrative data. For information on both aspects, see Haslett, Jones, Noble and Ballas (2010).

2. Methodology

2.1 Small area estimation – ELL / World Bank method

In this section, we present a brief overview of small area estimation, and the ELL (Elbers, Lanjouw and Lanjouw, 2003) method which is supported by the World Bank and has been used primarily for small area poverty estimation in developing countries. ELL has strong links to the economic rather than the statistical literature (e.g. Bramley, 1992; Bramley and Smart, 1996; Bramley and Lancaster, 1998) and was followed by various related publications (e.g. Elbers, Lanjouw and Leite, 2008).

2.1.1 Small area estimation

Small area estimation refers to a collection of statistical techniques designed for improving sample survey estimates through the use of auxiliary information (Rao, 2003). We begin with a target variable, denoted *Y*, for which we require estimates over a range of small subpopulations, usually corresponding to small geographical areas. Direct estimates of *Y* for each subpopulation are usually available from sample survey data, in which *Y* is measured directly on the sampled units (households or eligible children). Because the sample sizes within the subpopulations will typically be very small, these direct estimates will have large standard errors and hence will not be reliable. Some subpopulations may not even be sampled at all in the survey. Auxiliary information, denoted *X*, can be used under some circumstances to improve the estimates, giving lower standard errors for sampled areas and estimates even for unsampled areas.

Letting *X* represent additional variables that have been measured for the whole population, either by a census or via a GIS database. A relationship between *Y* and *X* of the form $Y = X\beta + u$ can be estimated using the survey data, for which both the target variable and the auxiliary variables are available. Here β represents the estimated regression coefficients giving the effect of the *X* variables on *Y*, and *u* is a random error term representing that part of *Y* that cannot be explained using the auxiliary information. If we assume that this relationship holds in the population as a whole, we can use it to predict *Y* for those units for which we have measured *X* but not *Y*. Small area estimates based on these predicted *Y* values will often have smaller standard errors than the direct estimates, even allowing for the uncertainty in the predicted values, because they are based on much larger samples. Thus the idea is to “borrow strength” from the much more detailed coverage of the census data to supplement the direct measurements of the survey.

2.1.2 Clustering

The units on which measurements have been made are often not independent, but are grouped naturally into clusters of similar units. When such structure exists in the population, the regression model above can be more explicitly written as

$$Y_{ij} = X_{ij}\beta + c_i + e_{ij} \tag{3.1}$$

where Y_{ij} represents the measurement on the *j*th unit in the *i*th cluster, c_i the error term held in common by the *i*th cluster, and e_{ij} the household-level error within the cluster. The relative importance of the two sources of error can be measured by their respective variances σ_c^2 and σ_e^2. Ghosh and Rao (1994) give an overview of how to obtain small area estimates, together with standard errors, for this model. Where individual-level rather than household-level data is used, an additional error term at within household is added. In the general explanation given below we focus on equation (3.1) in order to establish general principles useful for distinguishing the characteristics of variation at 'higher' and 'lower' levels. It is also possible, and indeed strongly advisable to add an error term at small area level, to check whether the variables included in the model are sufficient to rule out the requirement for small area level random effects.

We note that the auxiliary variables X_{ij} may be useful primarily in explaining the cluster-level variation, or the household-level variation. The more variation that is explained at a particular level, the smaller the respective error variance, σ_c^2 or σ_e^2. The estimate for a particular small area will typically be the average of the predicted *Y*s in that area. Because the standard error of a mean gets smaller as the sample size gets bigger, the contribution to the overall standard error of the variation at each level, household and cluster, depends on the sample size at that level. The number of households in a small area will typically be much larger than the number of clusters, so to get small standard errors it is of particular importance that, at the higher level, the unexplained cluster-level variance σ_c^2 should be small. (A parallel comment applies to any small-area level variance in comparison with the cluster level variance.)

Another important aspect of clustering is its effect on the estimation of the model since to account properly for the complexity of the survey design requires the use of specialized statistical routines (Skinner, Holt and Smith., 1989; Chambers and Skinner, 2003; Lehtonen and Pakhinen, 2004; Haslett, Isidro and Jones, 2010) to get a consistent estimate for the regression coefficient vector β (i.e. $\hat{\beta}$) and its variance $V_{\hat{\beta}}$.

For ELL, the size of the standard error depends on a number of factors. The poorer the fit of the model (3.1), in terms of small R^2, large σ_c^2 or σ_e^2, or a large $\sigma_c^2/(\sigma_c^2+\sigma_e^2)$ ratio, the more variation in the target variable will be unexplained and the greater will be the standard errors of the small area estimates. The population size and the sample size of the survey to which the model is fitted are two important factors.

The integrity of the estimates and standard errors depends on the fitted model being correct, in that it applies to the census population in the same way that it applied to the sample. This relies on good matching of survey and census to provide valid auxiliary information. Spurious relationships or artefacts which appear, statistically, to be true in the sample but do not hold in the population can be caused by fitting too many variables, or by choosing variables indiscriminately from a very large set of possibilities leading to severe underestimation of the standard error. Including small area level error effects in models is also crucial if standard errors are not to be underestimated, unless the variance of these effects is sufficiently small.

2.2 Spatial microsimulation: Synthetic reconstruction, spatial microsimulation and combinatorial optimisation methods

2.2.1 Aspatial microsimulation

There are microsimulation models that do not take geography into account. It can be argued these are "aspatial" or non-geographical. There is an extensive literature on this topic, beginning with the work of Orcutt (1957), and Orcutt, Greenberger, Korbel and Rivlin (1961). An extensive review is given in O'Donoghue (2001).

Aspatial microsimulation is a technique developed, particularly by economists, that has been widely used for over 50 years. The results of microsimulation models are also widely quoted in the UK and USA media when covering possible impact of government budget changes upon different types of households. The models have aimed to build large-scale data sets on the attributes of individuals or households (and/or on attributes of individual firms or organisations) and to analyse policy impacts based on these micro-units (e.g. Orcutt, Mertz and Quinke., 1986). By analysis at the level of the individual, family or household, they claim to provide the means of assessing variations in the distributional effects of different policies (Hancock and Sutherland, 1992; Mitton, Sutherland and Weeks, 2000). Microsimulation modelling frameworks have become accepted tools in the evaluation of economic and social policy, as well as analysis of tax-benefit options and other areas of public policy.

Statistics Canada (http://www.statcan.ca/english/spsd/) has produced several microsimulation models. One of these is the *Social Policy Simulation Database and Model* (SPSD/M) which was designed to analyse the financial interactions of governments and individuals in Canada and the cost implications and income redistributive effects of changes in the personal taxation and cash transfer system.

Generally, microsimulation models have wider application when they become *dynamic*, by updating once a microsimulation database is built. Among the first usable dynamic microsimulation models is DYNASIM (DYNAmic Simulation of Income Model; see Orcutt, Greenberger, Korbel and Rivlin, 1961; Wertheimer, Zedlewski, Anderson and Moore, 1986), which was the base for more sophisticated developments such as CORSIM (Cornell Microsimulation Model – Caldwell, Clarke and Keister, 1998) and DYNACAN (Dynamic Microsimulation Model for Canada). One of the descendants of DYNASIM was DYNASIM2, developed and maintained at the Urban Institute in Washington D.C. (Wertheimer *et al.,* 1986). Other relevant microsimulation models worldwide include DYNACAN in Canada, and DYNAMOD in Australia.

As the previous comments, and the extended commentary and list of references in Haslett, Jones, Noble and Ballas (2010), indicate, microsimulation modelling has been used extensively. Its widest application has been to assess the future effect of policy changes. Such assessments are complicated not only by choice of underlying model, but also by what should be a requirement to assess accuracy of predictions and the long term nature of many of the predictions. Historically, modelling accuracy has not been assessed. Where differences in consequences between scenarios has been marked, accuracy is perhaps not so important, but where differences are more subtle, accuracy measures may be crucial. Models used have often been decided a priori based on expert opinion, and this too has added to uncertainty in predictions, because the models have not been formally tested statistically. These aspects can be major complications for the effectiveness of microsimulation.

2.2.2 Spatial microsimulation

Microsimulation models become geographical or spatial when area-based information about the simulated entities is available (or estimated). In particular, geographical microsimulation can be defined here as a method to construct small area population microdata for one point in time and then to update these microdata. This definition of microsimulation is different from that used by economists involved in building statistical and mathematical microsimulation models, since these generally do not take geography into account. The focus in spatial microsimulation is to provide small area socio-economic information that can be used for the spatial analysis of policies, as well as the inter-household distributional effects.

Spatial microsimulation involves the creation of large-scale population microdata sets and the analysis of policy impacts at the micro-level (e.g. Ballas et al, 2007). Population microdata can be individual microdata that contain information on individuals; household microdata which may contain household information only and household microdata which may contain individual and household information.

This section discusses how various aspatial methods and techniques can be refined and applied in a geographical context in order to provide small area microdata.

Small area microdata can be built with the use of static spatial microsimulation methods. We can distinguish between the following types:

- Synthetic probabilistic reconstruction models, which involve the use of random sampling
- Reweighting probabilistic approaches, which typically reweight an existing national microdata set to fit a geographical area description on the basis of random sampling and optimisation techniques
- Reweighting deterministic approaches, which reweight a non geographical population microdata set to fit small area descriptions, but *without* the use of random sampling procedures

These approaches are discussed in detail in Haslett , Jones, Noble and Ballas (2010) along with information on how the method has been used as a dynamic model to project results to future dates for policy assessment purposes.

Results and outputs from dynamic microsimulation techniques hold great attraction for planners and policy makers, but some notes of caution are warranted. In addition to the complications of static microsimulation, in particular the risk of poorly specified and untested underlying models and no standard error estimates, there is the additional problem of projecting or predicting data. Again models are at the core in projection, and the time series available are almost inevitably short, making explicit model extrapolation and testing fraught. Sophistication in projection models is a commendable aim, but fitting (let alone testing) such models is far from simple, and the projections remain very dependent on the type of projection model chosen. There is also the temptation to ask quite reasonable policy and other questions that go beyond the ability of the method to answer, given the strong assumption inherent in its construction and fitting. Finally, such models are used to generate microdata which gives the superficial appearance of being a census, but is not. Producing multiple datasets for every scenario considered would at least have the advantage of allowing assessment of accuracy conditional on the model being correct, and consequently this multiple imputation approach is highly recommended.

2.3 Mass imputation

Mass imputation is a technique trialled and used by Statistics Canada (e.g Colledge, Johnson, Paré and Sande, 1978; Kovar and Whitridge, 1995) and Statistics Netherlands (Kooiman, 1998; de Waal, 2000) but has fallen out of favour at both institutions.

Mass imputation covers a wide range of techniques, or rather imputation models, with the common features that a high or very high percentage of the data is imputed. Not only a particular variable but a complete record for a respondent may be imputed. It has most often been applied to survey data, usually to supplement it where there is substantial item non-response, or to create a pseudo-census.

At Statistics Canada, mass imputation was first used in the context of two phase sampling of administrative records. An efficient design was used to select a first phase sample from which additional information was collected. As an alternative to using sample survey weighting, imputation was used for the missing parts of the non-sampled primary units to produce a complete, rectangular file. This technique is what Statistics Canada call mass imputation (Kovar and Whitridge, 1995 p. 413). Statistics Canada first applied mass imputation to its Census of Construction data (Colledge *et al*., 1978), where the imputation rate was approximately 70%. It has also used mass imputation for agricultural income tax data to produce balance sheet estimates, where farmers' records are missing for operational reasons rather than at random.

At Statistics Netherlands, mass imputation has been largely superseded by iterative reweighting methods that match survey data to a range of consistent tables, some from other surveys, some from administrative registers, without (as for deterministic, but different from probabilistic spatial microsimulation) creating a full pseudo-census. The Dutch call this technique (or perhaps more strictly the data generated from it) a "virtual census". Although sometimes quoted in support of mass imputation, Houbiers (2004, p. 56-57) actually notes:

> In principle, mass imputation offers a simple alternative to estimation by weighting to achieve numerical consistency between estimates from the [Social Statistical Database] SSD. By using some suitable imputation strategy, all missing fields in the SSD can be imputed. Tables can then simply be "counted" from the resulting complete data set. Although imputation models are better when more register information is available, these models are never sufficiently rich to account for all significant data patterns between sample and register data, and may easily lead to oddities in the estimates (see Kooiman 1998). Therefore, traditional estimation by weighting is favored over mass imputation at Statistics Netherlands.

The Dutch virtual census approach is possible because of extensive register data in Holland (as in Scandinavia); available registers and sizes at 2001 included the Population Register (16,000,000 records), the Jobs File of employees (6,500,000 records), the Fiscal Administration database (jobs: 7,200,000 records, and pensions and life insurance benefits: 2,700,000 records), Social Security Administration (2,000,000 records), and surveys included the Survey of Employment and Earnings (3,000,000 records – working hours, place of work) and the Labour Force Survey (2 years, 230,000 records: education, occupation and economic activity). Together these provide a very rich data source, but in many other countries such extensive information is not available, and this limits general utility of the virtual census method for Official Statistics.

Kovar and Whitridge (1995) make a number of insightful remarks about mass imputation and its use:

- For many imputation methods there is a corresponding weight adjustment: For simple random sampling using the sample mean for imputation is equivalent to the direct expansion estimator. Using the ratio estimator with auxiliary data to mass impute for subsampled variables is equivalent to using a ratio estimator.

- Nearest neighbour imputation is implicitly equivalent to an expansion estimator with variable weights corresponding to the number of times each sampled record is used as a donor.
- Weighting rather than mass imputation is recommended for more complicated statistics, such as variances, covariances and correlations.
- Mass imputation "has a place" where "quick, ad hoc estimates are needed, or where second-phase sample weights are difficult to calculate…as when information is missing for operational reasons".
- For non-random ignorable non-response, mass imputation by nearest neighbour methods may be preferable to weighting, since it makes more extensive use of auxiliary information and multivariate relationships, and may help attenuate bias.
- Mass imputation performs very poorly where there is non-ignorable non-response.
- The choice of imputation method is important.
- Significant bias can be introduced by variables that are not controlled in the imputation process.
- Imputed values should be flagged.
- Evaluation of the effect of mass imputation is critical.

Given these caveats, mass imputation has nevertheless more recently been under discussion and/or in use, in agriculture statistics (Fetter, 2001), at the Australian Bureau of Statistics (ABS, 2003), at Statistics Norway (Gåsemyr, Børke and Andersen, 2007), at INSEE in France (Brion, 2008), and in results given at various conferences (e.g. Kozak, 2005; Kroti, Black and Creel, 2005).

Regardless of whether multiple imputation is used to allow estimates of accuracy conditional on the model, the success of the mass imputation technique itself depends critically on the adequacy of the imputation model, both in terms of model type and the variables included in it. Simple hot-decking, i.e. using a single pass through the data and replacing missing records from imputation classes formed from cross-tabulations of the data, is not generally suitable for mass imputation. Simple reweighting is generally preferred because it removes a random element due to random record choice, it is computationally less intensive, and it has a large body of theory detailing its properties.

2.4 Associated techniques

Mass imputation is only one of a variety of imputation techniques. Others include multiple imputation, fractional imputation, various varieties of hot decking, deterministic and stochastic imputation. These methods are not mutually exclusive. A very useful reference remains Kovar and Whitridge (1995).

As part of the procedure, to impute for missing data, fractional imputation (Kim and Fuller, 2004) adds a fraction of a randomly chosen residual to a regression-based predictor, where the fraction is a function of independent variables in a regression. Although (unlike ELL) fractional imputation is usually applied to survey data, it has a connection to ELL where the bootstrap residuals used are also scaled or "unshrunk".

Inverse sampling (Hinkins, Oh and Scheuren, 1997; Rao, Scott and Benhin, 2003) rather than trying to recreate a complete census from survey data instead subsamples the survey data (perhaps many times) to produce a dataset (or datasets) that can be analysed as though they were simple random samples. A simple example is for a cluster sample with equal cluster sizes and sample size within sampled clusters, where each subsample would contain one unit selected randomly from each sampled cluster. This technique is intended to produce a ‘complete’ dataset, even for a survey.

Record linkage methods also combine data sources to create a single dataset in the absence of unique identifiers, but unlike the methods considered here the datasets are usually of similar size. Generalized regression (GREG) estimation methods are usually applied to sample survey data only, rather than used to combine datasets.

M-quantile estimation (Chambers and Tzavidis, 2006) has been suggested as an alternative to ELL, but has not been assessed in this study.

Reweighting of data, essentially using calibration, raking, or the iterative proportional fitting algorithm (which are essentially equivalent techniques) is an inherent part of post-data collection procedures for survey data in many government statistical agencies. The aim is to adjust and thus match different tables to be consistent, i.e. to have the same margins for the same variables (or to match survey totals with census ones where known by adjusting survey weights).

IPF and SPREE (Structure PREserving Estimation) methods are also inherently related to spatial microsimulation where survey microdata are chosen or rescaled to match census margins, and hence to loglinear models (see, for example, Noble, Haslett and Arnold, 2002; Haslett, Noble and Arnold, 2006), the generalised version of SPREE, GSPREE (Zhang and Chambers, 2004) and its extended version, ESPREE (Isidro, 2010).

There are also very close links to ecological inference and data fusion, except there is then no sample model linking Y and X. Recent developments in the ecological inference literature (see Steel, Beh and Chambers, 2004) have identified the considerable gains from having even a small amount of linked data, while researchers in data fusion are now (somewhat belatedly) coming to the same conclusion.

3 A unifying theoretical framework for ELL-type small-area estimation, spatial microsimulation, and mass imputation

ELL-type small-area estimation, spatial microsimulation and mass imputation are all techniques that use survey and partial census or administrative data to create a pseudo-census. While there are important, if not critical differences between them that affect their utility, these tend to hinge on the model chosen, how it is selected, fitted and tested, and whether the method (as currently used) provides estimated standard errors conditional on the model.

In this section we do not focus on these differences, although they do help explain why the three methods do not work equally well. Instead we consider why and in what ways the three methods are fundamentally similar.

We assume that the object of interest is a (possibly nonlinear) function of the complete census data, say $\varphi(C)$. In general, the operator $\varphi(.)$ will act on a target variable or variables Y contained in the census; for example in small area estimation $\varphi(.)$ will typically produce subpopulation means of Y – a linear function – whereas in small-area estimation of poverty “incidence” the object is the subpopulation proportion of Y (income or expenditure) values below a threshold – a nonlinear function. In some uses of spatial microsimulation, $\varphi(.)$ may involve the application of a simulation model to household - or individual- level data; this too can be regarded as a nonlinear function of the census observations.

All three methodologies have been developed to cope with situations in which the complete census data is unobserved. We write formally:

$$C = C_O + C_U$$

where C_O, C_U denote respectively the observed and unobserved portions of the full census data. In small-area estimation situations, for example, we often have complete census data for "auxiliary variables" *X*, but only partial, survey-derived, data for the target variable *Y*. In mass imputation and spatial microsimulation, some records may be more incomplete than others, and many census observations may be missing most of the auxiliary data in addition to the variable(s) of interest.

We seek to replace the target:

$$\varphi(C) = \varphi(C_O + C_U)$$

by an estimate:

$$\varphi(C^*) = \varphi(C_O + C_U^*)$$

in which the missing data C_U is replaced by a surrogate C_U^*. This surrogacy is or should be informed by a "model", i.e. a set of assumptions or a fitted statistical structure that tells us what to expect for C_U based on C_O.

In small-area estimation with an explicit linear model, C_U^* will be the expected value of C_U conditional on the observed data:

$$\varphi(C^*) = \varphi(C_O + E[C_U \mid C_O])$$

In poverty estimation, where incidence is a nonlinear function of income or expenditure, the ELL method is targeted at:

$$\varphi(C^*) = E[\varphi(C_O + C_U) \mid C_O]$$

In stochastic microsimulation, the actual households in an area for which complete data is unavailable are replaced by a set of households with complete data, chosen to match some of the characteristics of the actual households. This again implies a model, since it assumes that the characteristics matched are useful in predicting the variable(s) of interest. Again denoting the variable(s) of interest by *Y*, and the matching characteristics by *X*, the area-level summaries $\varphi(C)$ are approximated by draws from the distribution of:

$$\varphi(C^*) = \varphi(C_O + C_U \mid C_O)$$

If deterministic microsimulation is used (via iterative proportional fitting) or if multiple stochastic microsimulation estimates are averaged, the situation is then exactly the same as that for small-area estimation as detailed above.

In mass imputation, incomplete records in the census have their missing portions replaced using complete records that match on the non-missing portions. Here again we can regard C_U^* as a random draw from the distribution of $C_U \mid C_O$. If multiple imputations are used, these can be averaged to again give:

$$\varphi(C^*) = E[\varphi(C_O + C_U) \mid C_O]$$

Because some of the auxiliary data *X* may be missing, in addition to some of the *Y* values, this process is equivalent to a model-based small-area estimation in which noise has been added to some of the *X* variables, the amount of noise being determined by the amount of missingness in *X* and the size of the model errors in the imputation of the missing *X*s.

This framework forms the basis for the extensive simulation study presented in Haslett, Jones, Noble and Ballas (2010).

4. Conclusions

4.1 Simulations

The extensive simulations in Haslett, Jones, Noble and Ballas (2010) indicate underlying similarities between spatial microsimulation, small area estimation using the ELL method, and mass imputation, in line with the structural similarities from a more theoretical perspective apparent from Section 3 above. These similarities do not however extend to categorising all three methods together in terms of effectiveness, since in practice they tend to be used in different ways.

Mass imputation and spatial microsimulation have tended to be used with implicit, and perhaps too often untested, statistical models as their basis, with variables included in them being decided a priori or on the basis of the (often rather limited) number of variables available. In the case of spatial microsimulation, the information available has taken the form of various census cross-tabulations and variables, which implicitly define which effects are included in a loglinear model. For mass imputation the situation is even more opaque, as even the effect of the best imputation methods based on nearest neighbour techniques (even if made explicit) do not necessarily lead to a clearly specified statistical model.

Small area estimation, whether using ELL or not, has a longer history of specifying, fitting and testing explicit statistical models, and it is recommended that such specifying, fitting and testing provide a focus for further research in spatial microsimulation and mass imputation.

All three methods, under well-specified and fitted models without bias, are capable of producing reliable estimates, even where projections of data are required.

4.2 General conclusions: links and comparison of small area estimation (ELL), spatial microsimulation, and mass imputation

In practice, despite an underlying conceptual and theoretical similarity and that all are methods for ‘completing’ databases, there are both similarities and differences between spatial microsimulation, mass imputation, and small area estimation using the ELL method.

For all three techniques, the generally common intention (either as an interim or final output) is to produce a dataset (or datasets) which is rectangular without missing values, created by substitution of missing information using an implicit or explicit statistical model. The attraction of this approach is that, superficially at least, the pseudo-data provides a substitute for the unavailable data, though caution is clearly warranted for complex statistics required accurately by small area. The three methods can be considered as variations on a theme, under the unifying framework outlined in Section 3. This unification has a number of consequences:

- the structure of the underlying statistical model (e.g. linear or non-linear, with or without random effects) needs to be determined on strong theoretical grounds.

- the model needs to be fitted and tested, and should explain a substantial part of the variation in records not requiring imputation, so that inference to incomplete records can be properly justified.
- imputing residuals only, rather than entire variables, has major advantages in terms of utility of the pseudo-census(es), since it is better able to control bias especially where average values (for example for areas) are required.
- estimation of standard errors conditional on the fitted model is possible not only for ELL-type small area estimation (where it is routine), but also by a simple theoretical extension under the common framework of Section 3 to spatial microsimulation and mass imputation.

ELL-type small area estimation currently has the advantage over the other techniques of an explicit statistical model which is not only specified, but also fitted and tested. It is also able to provide estimated standard errors for its area-level averages.

One issue that deserves further discussion about ELL, however, is that ELL has smaller (sometimes much smaller) estimated standard errors than many of the small area methods of Rao (2003) and Longford (2005). This is not a direct result of the modelling, since all these small area methods first fit models to the survey data and test them, but instead due to the levels at which the error structure of models are fitted and to differences in the way available census data is integrated into the small area estimates.

ELL does not include a small-area-level error in its models. Instead it includes cluster (within area) and household (within cluster) error terms in linear models that may contain a comparatively large number of predictor variables, fitted separately to each survey stratum. While this strategy limits omitted variables (and hence the need for a small-area-level error term), it runs the counter-risk of overfitting models, since the number of candidate variables (including interactions) is often close to the number of observations within strata. Not all ELL-type models run similar risk of overfitting however or fail to consider area level random effects (see, for example, Jones, Haslett and Parajuli, 2006, where the small-area-level error has negligible effect on the standard error estimates of poverty).

Many of the models fitted to survey data by Rao (2003) and by Longford (2005) do not use census data at all or only census averages by small area, so their accuracy is determined by the fitted model (including any small-area-level errors) and limited (even if indirectly) by survey sample size. In comparison, for ELL every census observation can be and is predicted (multiple times, under the model that has been fitted to the survey data), based on the regression and added imputed residuals. ELL might be viewed as involving mass imputation (J. N. K. Rao - personal communication) but if so it is mass imputation of residuals only and usually for one variable (rather than a range of variables, as is more usual in mass imputation) under a comparatively well specified and tested model where the bulk of the predictor is fully based on a model, all predictor variables are available for all census observations and have been matched against their survey equivalents both in definition and in value, and where taken over small areas the expected value of the residual under the model is zero (which is a property that can be tested). So this is not mass imputation in the usual sense, since it is only for univariate residuals not multivariate observations, and, whatever it is called, it is comparatively small part of the prediction, especially after individual or household pseudo-census observations are aggregated to small area level. Nevertheless, ELL tends to produce much smaller estimated standard errors for the same small areas than mixed model methods that include small-area-level errors and do not integrate (or do not so fully integrate) known census information on key predictor variables. One explicit reason for the difference in standard errors is that the contribution of the estimated variance components in the ELL model (which are themselves similar if not identical to those estimated in a survey based small area method) are divided by the population size (e.g. number of clusters, or households) for estimated standard errors for small areas from ELL, rather than by the corresponding sample size as required when census predictions are not available or used. It is this stronger model assumption in ELL (which can be tested as part of the model fitting and only applies to the residuals anyway) together with the assumption that the model contains enough predictor variables that the small-area-level error is negligible, that gives ELL its markedly lower estimated standard errors than small area techniques not incorporating census data so directly. The point is not that ELL is wrong, but that it requires stronger assumptions, which must be rigorously tested as part of the model fitting process. Related issues are discussed in Haslett and Jones (2005a, 2005b), Jones and Haslett (2003), and Jones, Haslett and Parajuli (2006).

Unlike the other techniques, the subclass of deterministically reweighted spatial microsimulation methods do not necessarily produce a pseudo-census. If its weights for survey observations were integer, creating a pseudo-census using this method would involve only one additional step: simple replication of observations, so that the number of replications equalled each survey observation's weight. More often however, weights from deterministic spatial microsimulation are non-integer. Deterministically reweighted spatial microsimulation (unlike its probabilistic reweighted relation) is a genuine reweighting technique, based on use of IPF or its equivalent to calibrate to various census and other tables. In fact, deterministically reweighted spatial microsimulation is substantially different from the Dutch "virtual census" only in the methods used for choosing calibration variables (which are rather more implicit for deterministically reweighted spatial microsimulation) and in the possibility of using observations from outside the small area in spatial microsimulation.

Spatial microsimulation and mass imputation may impute complete or near complete records. In practice though, for both techniques at least some variables are available for all records, usually as aggregate counts by small area from census or administrative data sources, and these are used to inform an implicit imputation model, which is usually decided a priori rather than tested statistically before adoption. For mass imputation, the technique used (e.g. nearest neighbour imputation) may be set, but the model is still usually implicit. One intriguing possibility is that, where the model is implicit, its performance may be testable using the techniques used in data mining, e.g. cross-validation.

Small area estimation using the ELL method does not impute complete records, but instead usually imputes only one variable at a time under a mixed linear model. For ELL it is only the residuals from the random components in the mixed model that are imputed; most of the structure in the imputation model is contained in a regression equation.

In summary, even though all three methods, spatial microsimulation, mass imputation, and small area estimation via ELL, show strong structural similarities, this does not mean that deficiencies in one are necessarily deficiencies in another. Of the three methods, the underlying model used for imputation is explicit only in ELL, and consequently ELL can really be considered the best of the techniques given sound model fitting and testing. Adding such explicit fitting and testing to spatial microsimulation and mass imputation would improve both techniques, without being theoretically burdensome. It would also allow their accuracy to be better assessed, as is already done for small area estimation using ELL, by creating multiple pseudo-censuses and estimating standard errors under the specified and tested model. From these points of view, rapid improvements to the accuracy and assessment of accuracy of spatial microsimulation models in particular should be plausible under the theoretical framework for all three techniques outlined in Section 3.

Guidelines on when it is advisable to use these three techniques, and how best to design a sample survey when their use is intended are given in Haslett, Jones, Noble and Ballas (2010).

Acknowledgements

This research project was funded under the New Zealand Official Statistics Research (OSR) Programme 2008/9 through Statistics New Zealand. The opinions, findings, recommendations and conclusions expressed in this report are those of the author(s), do not necessarily represent Statistics New Zealand and should not be reported as those of Statistics New Zealand. The department takes no responsibility for any omissions or errors in the information contained here.

References

Australian Bureau of Statistics (2003). “The measurement strategy for register shocks?”, *Australian Bureau of Statistics Methodological News*, September 2003. http://www.abs.gov.au/AUSSTATS/abs@.nsf/Previousproducts/1504.0Main%20Features900Sep%202003?opendocument&tabname=Summary&prodno=1504.0&issue=Sep%202003&num=&view=

Ballas, D.**,** Kingston, R., Stillwell, J., and Jin, J. (2007). “Building a spatial microsimulation-based planning support system for local policy making”, *Environment and Planning A*, 39(10), 2482 - 2499.

Bramley, G (1992). “Homeownership affordability, in England”, *Housing Policy Debate* 3(3), 143-182.

Bramley, G. and Lancaster, S. (1998). “Modelling local and small-area income distributions in Scotland”, *Environment and Planning C*, 16,68 l-706.

Bramley, G. and Smart, G. (1996). “Modelling local income distributions in Britain”, *Regional Studies,* 30(3), 239-255.

Brion. P. (2008). “The future system of French structural business statistics: the role of the estimates”, paper presented to the *United Nations Statistical Commission / Economic Commission for Europe: Conference of European Statisticians*, Work session of statistical data editing, Vienna, 21-23 April 2008. http://www.unece.org/stats/documents/2008/04/sde/wp.13.e.pdf

Caldwell, S. B., Clarke, G. P. and Keister, L. A. (1998). “Modelling regional changes in US household income and wealth: a research agenda”, *Environment and Planning C: Government and Policy*, *16*, 707-722.

Chambers, R. and Tzavidis, N. (2006). “M-quantile Models for Small Area Estimation”, *Biometrika*, Vol. 93, 255-268.

Chambers, R. L and Skinner, C. J. (eds.) (2003). *Analysis of Survey Data*. Wiley, New York.

Colledge, M. J., Johnson, J. H., Paré, R. and Sande, I. G. (1978). “Large scale imputation of survey data”, *Survey Methodology*, 4, 203-224.

Elbers, C., Lanjouw, J. and Lanjouw, P. (2003). “Micro-level estimation of poverty and inequality”, *Econometrica*, 71, 355-364.

Elbers, C., Lanjouw, P. and Leite, P. G. (2008). *Brazil within Brazil: testing the Poverty Map Methodology in Minas Gerais*, Policy Research Working Paper 4513, World

Bank Development Research Group - Poverty Team, February 2008, World Bank, New York.

Fetter, M. (2001). “Mass imputation of agricultural economic data missing by design: a simulation study of two regression based techniques”, *Federal Conference on Survey Methodology*. http://www.fcsm.gov/01papers/Fetter.pdf.

Gåsemyr, S., Børke, S. and Andersen, M. Q. (2007). “A strategy to increase use of administrative data in business statistics”, *Seminar of Registers in Statistics – Methodology and Quality*, Helsinki, May 21-23 2007. http://www.stat.fi/registerseminar/session3_gosemyr_pres.pdf

Ghosh, M. and Rao, J.K.N. (1994). “Small area estimation: an appraisal”, *Statistical Science*, 9, 55-93.

Hancock, R. and Sutherland, H. (eds.) (1992). *Microsimulation models for public policy analysis: new frontiers*, Suntory-Toyota International Centre for Economics and Related Disciplines – LSE, London.

Haslett, S., Isidro, M. and Jones, G. (2010). “Comparison of survey regression techniques in the context of small area estimation of poverty”, *Survey Methodology*, Vol 36, No 2 (in press).

Haslett, S. and Jones, G. (2005a). *Estimation of Local Poverty in the Philippines*, Philippines National Statistics Co-ordination Board / World Bank, November 2005.

Haslett, S. and Jones, G. (2005b). “Small area estimation using surveys and censuses: some practical and statistical issues”, *Statistics in Transition*, 7(3), 541-555.

Haslett, S., Jones, G. Noble, A.and Ballas, D. (2010). “More for less? Using existing statistical modeling to combine existing data sources to produce sounder, more detailed, and less expensive Official Statistics”, *Official Statistics Report Series*, Vol 3, http://www.statisphere.govt.nz/official-statistics-research/series/2010/page1.aspx ISSN 1177-5017, ISBN 978-0-478-31520-2, 2010, 75 pages.

Haslett, S., Noble, A. and Arnold, G. (2006). “Erratum to: ‘Small Area Estimation via Generalized Linear Models’, Journal of Official Statistics, 18(1), 2002, 45-60”, *Journal of Official Statistics*, 22(1), 159-160.

Hinkins, S., Oh, H. L. and Scheuren, F. (1997). “Inverse sampling design algorithms”, *Survey Methodology*, 23, 11-21.

Houbiers, M. (2004). “Towards a social statistical database and unified estimates at Statistics Netherlands”, Journal of Official Statistics, 20(1), 55-75.

Isidro, M. (2009). *Intercensal updating of small area estimates*, unpublished PhD thesis, Massey University, New Zealand.

Jones, G. and Haslett, S. (2003). *Local Estimation of Poverty and Malnutrition in Bangladesh*. Bangladesh Bureau of Statistics and United Nations World Food Programme.

Jones, G., Haslett, S. and Parajuli, D. (2006). *Small area estimation of poverty, caloric intake and malnutrition in Nepal*, September 2006, Published by the World Food Programme, the World Bank and the Nepal Central Bureau of Statistics, 218 pages, ISBN 999337018-5.

Kim, J. K. and Fuller, W. A. (2004). “Fractional hot deck imputation”, *Biometrika* 91, 547-557.

Kooiman, P. (1998). “Mass imputation: Why not!?”, Research paper, BPA-no. 8792-98-RSM, Statistics Netherlands, Voorburg, 1998 [in Dutch].

Kovar, J. G. and Whitridge, P. J. (1995). “Imputation of business survey data”, Chapter 22 in *Business Survey Methods*, eds. Cox, B., Binder, D. A., Nanjamma Chinnappa, B., Christianson, A., College, M. J., Kott, P. S., John Wiley and Sons, New York.

Kozak, R. (2005). “The Banff system for automated editing and imputation”, *SSC Annual Meeting: Proceedings of the Survey Methods Section*, 1-10.

Krotki, K., Black, S., Creel, D. (2005). “Mass imputation”, ASA Section of Survey Research Methods, 3266-3269. http://www.amstat.org/sections/srms/proceedings/y2005/Files/JSM2005-000931.pdf

Lehtonen, R. and Pakhinen, E. (2004). *Practical Methods for Design and Analysis of Complex Sample Surveys*, 2nd Edition, Wiley, New York.

Longford, N. T. (2005). *Missing Data and Small Area Estimation*, Springer Verlag, New York.

Mitton, L., Sutherland, H. and Weeks, M. (eds.) (2000). *Microsimulation Modelling for Policy Analysis: Challenges and Innovations*, Cambridge University Press, Cambridge.

Noble, A., Haslett, S. and Arnold, G. (2002). “Small Area Estimation via Generalized Linear Models”, *Journal of Official Statistics*, 18(1), 45-60.

O’Donoghue C. (2001). “Dynamic Microsimulation: A Methodological Survey”, *Brazilian Electronic Journal of Economics*, 4(2) [on-line journal], available from: http://www.beje.decon.ufpe.br/v4n2/cathal.htm

Orcutt, G. H. (1957), “A new type of socio-economic system”, *The Review of Economics and Statistics*, 39, 116-123.

Orcutt, G. H., Greenberger, M., Korbel, J. and Rivlin, A. (1961). *Microanalysis of Socioeconomic Systems: A Simulation Study*, Harper and Row, New York.

Orcutt, G. H., Mertz, J., and Quinke, H. (eds.) (1986). *Microanalytic Simulation Models to Support Social and Financial Policy*, North-Holland, Amsterdam.

Rao, J. N. K. (2003). *Small Area Estimation*. Wiley, New York.

Rao, J. N. K., Scott, A. J. and Benhin, (2002). “Analysis of complex survey data using inverse sampling”, *Proceedings of Statistics Canada Symposium 2002: Modelling Data for Social and Economic Research*.

Rao, J. N. K., Scott, A. J. and Benhin (2003). “Undoing complex survey data structures: some theory and applications of inverse sampling”, *Survey Methodology*, 29(2), 107-128.

Skinner, C. J., Holt, D. and Smith, T. M. F. (eds.) (1989). *Analysis of Complex Survey Data*. John Wiley and Sons, New York.

Steel, D.G., Beh, E. J. and Chambers, R.L. (2004). “The information in aggregate data. In Ecological Inference: New Methodological Strategies”. (eds. G. King, O. Rosen and M. Tanner). Cambridge University Press: Cambridge.

de Waal, T. (2000). “A brief overview of imputation methods applied at Statistics Netherlands”, *Netherlands Official Statistics*, 15, Autumn 2000, 23-27.

Wertheimer II, R., Zedlewski, S. R., Anderson, J. and Moore, K. (1986). “DYNASIM in comparison with other microsimulation models”, in G. H. Orcutt, J. Mertz, H. Quinke (eds.) *Microanalytic Simulation Models to Support Social and Financial Policy*, North-Holland, Amsterdam, 187-206.

Zhang, Li-C. and Chambers, R.L. (2004). “Small area estimates for cross-classifications”, *Journal of the Royal Statistical Society*, Series B 66, 479-496.